AthenaMPI

A Multi-Node Version of ATLAS's Athena Framework, Using MPI (Message Passing Interface)
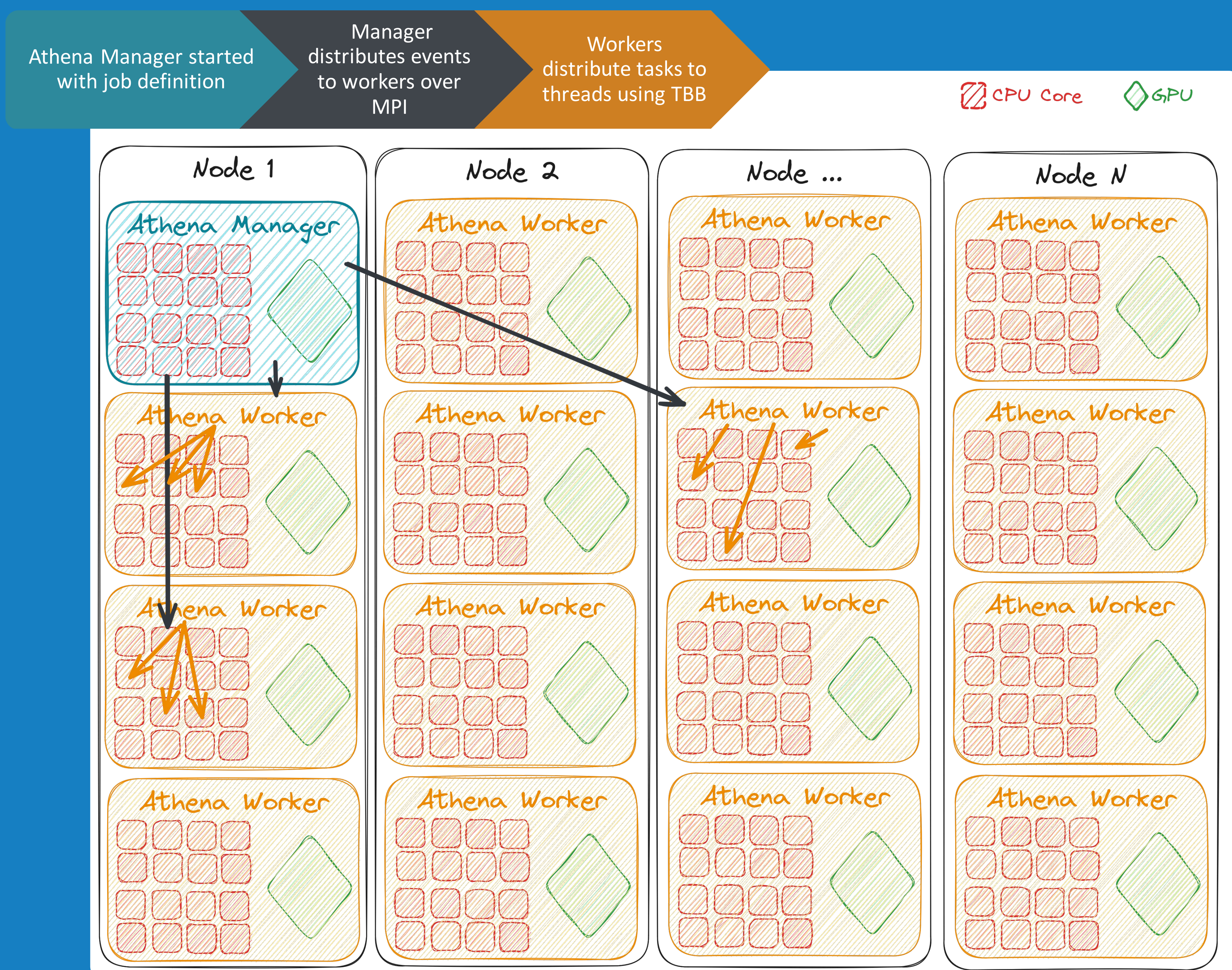
What is Athena?

- ⊙ Framework used for virtually all ATLAS computing
  - ⊙ From simulation to reconstruction to (some) final analysis
- ⊙ Started as a single-process, single-threaded framework
  - ⊙ Multi-process (forking) architecture added to save memory
  - ⊙ Later extended to full multi-threaded architecture
- ⊙ Remains a single-node framework

Multi-Node Capability

- ⊙ Traditionally used single-node jobs on grid
- ⊙ With HPCs, need to use multi-node jobs
  - ⊙ Don't want nodes sitting idle
  - ⊙ Non-trivial initialization time ⟶ Minimize number of initializations
- ∴ Need multi-node capability within Athena

Architecture

Athena Manager started with job definition → Manager distributes events to workers over MPI → Workers distribute tasks to threads using TBB

CPU Core   GPU

Node 1 — Athena Manager, Athena Worker, Athena Worker, Athena Worker
Node 2 — Athena Worker, Athena Worker, Athena Worker, Athena Worker
Node ... — Athena Worker, Athena Worker, Athena Worker, Athena Worker
Node N — Athena Worker, Athena Worker, Athena Worker, Athena Worker

Timeline and Performance

Raythena
- Ray wrapper around single-node Athena
- Many moving parts
- Shown at CHEP 2019

HPXGaudi
- Evaluated HPX and integrated in Gaudi
- Did not fit well with existing scheduler
- Scaling worse than expected
- Shown at CHEP 2023

AthenaMPI
- Re-evaluated technology choice – Switched to MPI
- Near ideal scaling
- Now integrated in Athena

ATLAS Simulation Preliminary
√s = 13.6 TeV
tt̄ Full Simulation with Geant4
256 × n_{nodes} Events

Execution Throughput [evts / sec] — Execution Throughput per Node [evts / sec]
Perfect Scaling
Nodes (Perlmutter CPU)

Paolo Calafiura, Julien Esseiva, Xiangyang Ju, Charles Leggett, *Beojan Stanislaus*, Vakho Tsulaia
Lawrence Berkeley National Laboratory

BERKELEY LAB
Bringing Science Solutions to the World

ATLAS EXPERIMENT